OIM 454
Individual Assignment 3

Data & Context

Please download the **Beach Attendance.csv** dataset from Canvas.

This file contains data regarding daily attendance figures to a set of beaches in a fictional county. Each row corresponds to a beach on a given day.

Below are the column definitions:

- Observation ID: unique identifier for each day-beach observation
- Beach ID: unique identifier for each beach
- Daily attendance: how many people attended the beach on that day
- Month
- Latitude: latitude of beach
- Longitude: longitude of beach
- Is Holiday: 1 if the day is a holiday, 0 if not
- Is Weekend: 1 if the day is a weekend, 0 if not
- Rain Index: continuous numerical variable measuring the amount of rain that day, 0 being the least
- AverageTemperature: average temperature of that day in Fahrenheit
- Wave Action: continuous numerical variable measuring the degree of wave action that day, 0 being the least

The county has a part-time staff of beach taggers, who go around and ensure that people pay a small fee to use the beach. These employees can only work a certain number of days a year. The county wants to deploy the beach taggers on the highest attendance days to generate the most revenue. To do this, they want to be able to predict when and on which day the most people will show up. Your job is to run a multiple linear regression on this data using R to predict each beach's attendance on a given day. You will evaluate how good this model is at identifying high attendance days using a lift chart.

Instructions

Open RStudio and create an R Markdown file. Save the .rmd file, ensuring that you save it in a file location that you can easily find again.

You will perform operations on this data using R. These operations must be performed in the order that they are specified in these instructions. Each of the following items corresponds to a separate code chunk that you must create in the R Markdown file. Make sure to run each code chunk after you write it to ensure that there are no errors. Be careful about typos!

If you have not installed the caret, ggplot2, and dplyr libraries yet, do so according to the instructions given in class before beginning this portion of the assignment.

Item 1: Load the caret, ggplot2, and dplyr libraries. Put this into a code chunk.

Item 2: Import the Beach Attendance.csv file into an R data frame called "beachattendance". Put this into a code chunk. HINT: You will need the full file path of the .csv file to do this. Consult the slides and/or the class recordings if you are unsure as to how to do this.

OIM 454
Individual Assignment 3

Item 3: Partition the data so that 70% goes into training, 15% goes into validation, and 15% goes into test.

Item 4: Run a multiple linear regression that predicts the daily attendance of a beach using all the usable predictor variables available in the dataset. (HINT: Do not use Observation ID and Beach ID as predictor variables.)

Item 5: What variables are positively correlated with daily attendance? What variables are negatively correlated with daily attendance?

Item 6: What is the R-squared on the training data?

Item 7. Use your trained linear regression model to produce predictions on the validation and test data.

Item 8. Produce evaluation metrics for the predictions made on both the validation and test datasets. What is the mean absolute deviation on the test data? What does this metric mean?

Item 9. Is this model overfit? Support your answer with specific evaluation metrics.

*Items 10-13: **OPTIONAL** **EXTRA CREDIT** as this was not covered in class. Feel free to supplement what was demonstrated in class by consulting the skeleton code as well as class slides to explain concepts. **+1 point for correct completion.***

> *Item 10. Prepare a dataframe for a cumulative gains chart using the test data predictions.*
>
> *Item 11. Calculate the lift metric for the test data predictions.*
>
> *Item 12. Plot the lift chart.*
>
> *Item 13. Use the lift chart to describe how the model performs on the test data as compared to the naïve benchmark. Do you think using this model will help the county maximize revenue from beach fees?*

Knit and export your R Markdown file as an HTML file. Upload the HTML file to Canvas to complete the assignment.

If you are unable to knit your .rmd file due to errors, make sure that you go back and test your code chunks individually. If you are ultimately unable to figure out how to solve these errors, save the .rmd file and upload that instead of the HTML file for partial credit.